DOCUMENT RESUME

ED 276 769 TM 860 720

AUTHOR Webster, William J.
TITLE Five Years of Teacher Testing: A Retrospective Analysis.
PUB DATE Apr 84
NOTE 34p.; Paper presented at the Annual Meeting of the American Educational Research Association (65th, New Orleans, LA, April 23-27, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; Beginning Teachers; Elementary Secondary Education; Essay Tests; *Evaluation Criteria; Evaluation Methods; Interviews; *Occupational Tests; Teacher Effectiveness; *Teacher Evaluation; *Teacher Selection; Test Construction; *Test Validity
IDENTIFIERS *Dallas Independent School District TX; National Teacher Examinations; *Wesman Personnel Classification Test

ABSTRACT

In 1977, the Dallas, Texas Independent School District modified its teacher selection criteria. In addition to the previously used evaluation criteria involving transcripts, references and evaluations, three new measures were employed: (1) a test of verbal and quantitative ability, the Wesman Personnel Classification Test (WPCT), which was later replaced by the locally developed Personnel Services Test; (2) the Personnel Services Department Essay Test; and (3) verbal response to a structured interview. These measures were intended to replace the National Teacher Examinations (NTE), because most applicants had not taken the NTE before the school year began. A series of studies was conducted over four years to validate the WPCT. Results showed a high correlation between the WPCT and the NTE, supporting the WPCT's concurrent validity. The Class Average Residualized Composite Score, an estimate of student performance in a teacher's classroom, was associated with a number of the measures. The WPCT predicted student achievement on the Iowa Tests of Educational Development at the secondary level. It was also found that WPCT scores were more powerful predictors of classroom performance than interviews or essays. Further study, however, indicated high reliability of the interview, which was a very important selection criterion. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED276769

FIVE YEARS OF TEACHER TESTING: A RETROSPECTIVE ANALYSIS[1]

William J. Webster
Dallas Independent School District

In the midst of a prevailing mood of discontent over public education throughout the nation and a growing perception reflected in the national media that much of this discontent was being fostered by so-called "incompetent" teachers (e.g., Ford, 1979; Gallup Poll, 1979; Lyons, 1978; U.S. News and World Report, 1978), the Dallas Independent School District (DISD) began a teacher selection program that was designed ultimately to upgrade the quality of public school education available to the children of Dallas. Since much of the output of an educational system seemed logically to be related to the quality of the teachers in that system, it was reasoned that one of the most rapid ways to increase the quality of that system was to increase the quality of its teachers.[2] Beginning in the fall of 1977, the DISD modified its teacher selection procedures to include three quantifiable criteria in addition to the traditional criteria that had, for many years, been used

---

[1] Paper presented at the annual meeting of the American Educational Research Association, New Orleans, Louisiana, April 23-27, 1984.

[2] DISD staff were apparently not alone in this thought. At the time the DISD proposed its new teacher selection system, ten states already required some form of teacher testing and an additional seven were considering such a requirement (Northern, 1980). By 1983, eighteen states required some form of teacher testing and another eleven were considering such a requirement (Feistritzer, 1983).

TM 860 720

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

W. Webster

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

in hiring teachers.[3] This resulted in the following eleven criteria being used in the teacher selection procedure:

Traditional Criteria

1. An analysis of the employment application.
2. An analysis of the applicant's certification.
3. An Analysis of the applicant's transcript.
4. An evaluation by the applicant's sending principal.
5. An evaluation by the applicant's cooperating teacher (if recent graduate).
6. An evaluation by the applicant's college supervisor (if recent graduate).
7. Recommendation by a District tri-ethnic screening committee.
8. An analysis of the applicant's personal and professional references.

New Criteria

9. A test of verbal and quantitative ability.
10. An essay response to a structured questionnaire.
11. Verbal response to a structured interview.

It is important to note that some operationalization of criteria 9, 10, and 11 were in place prior to the implementation of the new system; however, District decision-makers required more timely, reliable, and valid data than were available under the traditional system. This paper details the specifications and validation of the new teacher selection system and then examines, in retrospect, the effects of that system upon the improving quality of education available to students of the DISD.

Characteristics Of The System

Prior to November, 1977, DISD policy called for the use of National Teacher Examinations (NTE) scores in selecting prospective teachers; that is, NTE scores were required of teacher applicants before they

[3] With an annual turnover of between 400 and 600 teachers out of a 6,600 teacher work force, it was reasoned that a significant improvement could be made without experiencing the problems inherent in testing already-established teachers.

could be officially considered for employment. Unfortunately, only twenty-five to thirty-five percent of the applicants had normally taken the NTE by the time classes opened in August. The next prescribed date of administration of the NTE was generally in November; thus, four options were open to the District:

1. drop test scores as a prerequisite for employment
2. consider only twenty-five to thirty-five percent of the available applicant pool
3. require applicants to take the NTE after they had been employed and face the probability of unfortunate post-employment reanalysis
4. find a replacement test for the NTE

The decision was made to find a replacement test for the NTE; however, the test that was required was to be much different than the NTE. First, while the NTE was administered only by professionals commissioned by its publishing firm, the new test had to be administered by DISD professionals. Second, while the NTE in its shortest form (Common Examinations) had an administration time of three hours and fifteen minutes, the new test had to be capable of administration in one-half hour or less. This requirement precluded the use of an achievement test.

The test that was adopted for trial was the Wesman Personnel Classification Test (Psychological Corporation, 1965). The Wesman (WPCT) is a short (28 minutes) test of verbal and quantitative ability.[4]

---

[4] The quantitative ability section of the WPCT requires only minimal quantitative ability. Since the District was experiencing some discomfort over the failure of some personnel to perform minimal mathematical calculations, it was decided to include this section also.

It is not, nor was it intended to be, a teacher examination; that is, the WPCT makes no pretense of testing the substantive knowledge presumed necessary for good teaching. It was not designed to identify, by means of high scores, persons who would make excellent teachers. It was assumed, however, that persons who scored very low on the WPCT would be expected to encounter more-than-average difficulty in a profession that depends so much on one's ability to communicate. In short, it seemed logical that successful teachers should be minimally competent in acquiring, remembering, and transmitting knowledge. Levin's (1970) work on cost-effectiveness in teacher selection added additional weight to the choice of verbal ability as one variable in the selection system.

The WPCT subsequently became obsolete because of the inability of the District to keep the test secure. Within two years, evidence appeared that some local college and/or university students had the test.[5] As a result, in 1980 the DISD developed its own test of verbal and quantitative ability that could be kept secure. This test, which was designed to measure substantially the same kinds of skills that were measured by the WPCT, was entitled the Personnel Services Test (PST). The PST was designed according to the following guidelines:[6]

1. The verbal section of the test was to measure breadth of academic knowledge, the content domain of which is defined as informational elements in academic subjects covered in senior high school textbooks.

---

[5] Under normal circumstances there is a substantial correlation between the test score, the essay score, and the interview score. When this trend did not occur for students from one local college, subsequent investigation suggested that test security had been breached.

[6] Details about the development of the PCT are specified in a report by Carter and Taite (SP80-401-31-05).

2. The math section of the test was to be comprised of items on arithmetic and mathematical operations that were used by teachers on the job.

3. Item format for the verbal section of the test was to be single-entry, multiple-choice analogies that required minimal cognitive processing in determining the analogous relationships. The number of options on each item was to be four.

4. Duration of test administration was not to exceed 40 minutes for both sections.

5. Content coverage of the verbal subtest was to be partitioned equally into the following subject areas:

   English and Literature
   Social Studies (American History, Government, Economics, Geography, Psychology, and Current Events within the last decade)
   Science (Biology, Chemistry, and Physics)
   Miscellaneous (Sports, Music, Art, Theater Arts, Philosophy, Classics, Business, etc.).

Two other major changes in the teacher selection system involved the development of a standard essay test for applicants entitled the Personnel Services Department Essay Test (PSDET) and the development of a standardized interview procedure. The next section of this paper outlines the validation strategy for each of these new components as well as an examination of the validity of some of the traditional components in choosing successful teachers.[7]

The Validation Process

In attempting to validate the teacher selection system, a series of studies were designed over a period of four years. Several concerns became immediately apparent and were therefore the subject of early investigation. First, it was necessary to establish some measure of

---

[7] Details about the development of the PSDET are specified in a report by Taite (SP80-404-31-05). An evaluation of the interview process is reported by Carter and Berry (SP80-403-31-05).

concurrent validity for the WPCT. The selected approach involved correlating the WPCT with the Common Examinations of the National Teacher Examinations (NTE). The NTE are the most commonly used teacher examinations and, at that time, had recently withstood a legal challenge on the basis of content validity in the State of South Carolina (15 FEP Cases 1196 (1977); 38 CCh S. Ct. Bull, p. B559 (1978).) Second, it was necessary to determine how well performance on the essay exercise, the scored interview, and the WPCT predicted quality of instruction in the classroom. Quality of instruction was defined both according to the tenets of a renowned theory of instruction (Gagne, 1974) and, ultimately, by student gain-scores on nationally normed achievement tests.

The two aforementioned concerns were investigated through a series of two studies. The first study investigated the relationship between the NTE, the WPCT, and the personnel departmental interview and examined the contribution of each of these to the prediction of instructional quality (Carter, Casey, and Taite, IR78-283-31-12). Five instruments were used in this study. The WPCT was administered to 535 first-year classroom teachers, 288 of whom also had taken the Common Examinations of the NTE. The 535 teachers were also evaluated at the conclusion of their first year of teaching by their supervising principal. The evaluations were patterned after the District's official teacher evaluation system, except that principals were encouraged to be quite discriminating in their judgments and were assured that their evaluations were unofficial and would not be entered in personal files. As a result, the evaluations were subject to the usual lack of reliability of principal evaluations but not to the "halo" effect

usually associated with this type of evaluation. Each of the 535 teachers also had a structured interview score.

In addition, the students of the 31 secondary teachers included in the sample completed a measure of teacher instructional behavior that reflected Gagne's (1974) theory of pedagogy. The scale was composed of 31 items dealing with frequency of instructional events, four "halo" items, one interest-level item, and an overall assessment item. Students completed the scale in their teacher's absence. Coefficient Alpha estimates of the reliability of the overall student ratings was .91. Each of the 31 secondary teachers completed an identical measure to that completed by the students except that it was worded in the first person and also elicited information about relative class size and level of discipline in the classroom.

Seven instruments were used in the second study. Scores were available for 152 randomly selected first- and second-year teachers, stratified by level and subject area, on the WPCT. Forty-one randomly sampled first-year teachers were administered the aforementioned essay test. The inter-rater reliability on the test as measured by an unbiased intraclass coefficient of reliability was .88. Pre- and posttest scores (Spring to Spring) were available on every subtest of the Iowa Tests of Basic Skills (ITBS) for 2357 elementary students and 1841 middle school students who were associated with the teachers in the sample during the 1978-79 school year. Pre- and posttest scores (Spring to Spring) were available on every subtest of the Iowa Tests of Educational Development (ITED) for 1534 secondary students who were associated with the teachers in the sample during the 1978-79 school year.

To measure quality of instruction, ten observations were made on each classroom teacher utilizing an observation rating scale based on Gagne's (1974) theory of instruction. Separate scales were constructed for elementary and secondary teachers. Unbiased estimates of intraclass coefficients of reliability were .81 for the elementary scale and .80 for the secondary scale. A student rating scale, identical to the one described in the first phase of the study, was completed by 2493 middle and secondary school students associated with 70 secondary teachers.

The 535 first-year teacher scores on the WPCT were used to norm the test for the Dallas Independent School District (DISD) teacher population (Carter, Casey, and Taite, IR78-236-31-05). Since those results are of little interest to anyone but the District's Personnel Services Department and the press, they are not reported here. It should be noted, however, given the undue amount of bad publicity that Dallas teachers suffered, that those who had NTE scores scored, on the average, above the mean for their peers on the NTE.

Table 1 presents the Pearson zero order correlations among the WPCT verbal score, the WPCT math score, WPCT total score, the NTE common score, the interview score, and the principal's evaluation for all teachers in the sample with complete data (N=288).

See Table 1

Examination of the data presented in Table 1 suggests that the WPCT and the common examinations of the NTE are highly correlated. This fact provides some evidence for the concurrent validity of the Wesman. No other significant relations are apparent.

Table 1

Correlations among WPCT, NTE, Principals' Evaluation, and Interview Score

	WPCT-Verbal	WPCT-Math	WPCT-Total	NTE Common	Interview Score
Principals' Evaluation	.07	.04	.06	.03	.16
WPCT - Verbal		.56*	.93*	.81*	.02
WPCT - Math			.82*	.57*	-.05
WPCT - Total				.80*	-.01
NTE Common					.03

* $p \leq .001$

Table 2 presents the results of analyses of data pertaining to the 31 secondary teachers in the sample. Once again, Pearson zero order correlations are reported. Variables included in the analysis were WPCT verbal score, WPCT math score, WPCT total score, class size, teacher perception of discipline problems in the classroom, principals' evaluation of the teacher, verbal interview score, and a Class Average Residualized Composite Score (CARCS) for each teacher. All of the aforementioned variables except the CARCS are self-explanatory. The CARCS score was calculated as follows:

1. All behavior-related items and the interest and overall items on the student rating scale were residualized. That is, the effects of course subject matter, expected grade, the halo items, and student/teacher ethnic mix differential were statistically removed from the items.

2. A multiple regression of residualized behavior-related and interest items on the student residualized overall rating item was performed and resulting beta weights obtained.

3. The products of residualized items and their corresponding beta weights were added for each student.

4. An average was calculated by dividing the student sums by the number of students participating in the survey in each class. The average is the Class Average Residualized Composite Score (CARCS).

See Table 2

Although the sample is considerably smaller than the sample utilized in Table 1 (288 to 31), the correlations among the scores on the two standardized tests are quite comparable (.80, .82). In addition, correlations between the CARCS, a reliable estimate of classroom performance, and each of the standardized tests were significant. An interesting result was the relationship between discipline level and interview score. The -.42 suggests that the higher the interview score, the worse the level of classroom discipline. This finding would appear to be support for the time-worn axiom that "nice guys finish last." In addition, class size was consistently related to CARCS score, although the result was somewhat mitigated by the fact that lower scoring teachers tended to be assigned larger classes. It is important to note that both the NTE and the WPCT were significant predictors of CARCS scores with the effects of class size, discipline level, and interview scores partialled out (NTE = .38, WPCT = .40, both significant with $p \leq .05$). No other variables investigated in the study, with the exception of interview score being negatively related to discipline level, enjoyed the same degree of success.

In short, if CARCS is used as a measure of effective teaching, a multiple regression equation including WPCT total score, interview

Table 2

Correlations among NTE, WPCT, Class Size, Discipline Level, Principals' Evaluation, Interview Score, and CARCS for Secondary Teachers

	WPCT-Verbal	WPCT-Math	WPCT-Total	Class Size	Discipline Level	Principals' Overall Evaluation	Interview Score	CARCS
NTE Common	.81***	.63***	.82***	-.30	-.07	.04	.13	.47**
WPCT - Verbal		.57***	.93***	-.30	-.18	-.05	.18	.47**
WPCT - Math			.83***	-.26	-.17	-.13	.19	.37*
WPCT - Total				-.32*	-.20	-.09	.20	.48**
Class Size					-.07	.22	.16	-.47**
Discipline Level						.03	-.42**	-.01
Principals' Overall Evaluation							.14	.00
Interview Score								-.04

* $p \leq .05$, ** $p \leq .01$, ** $p \leq .001$

score, class size, and discipline level accounts for 34.8% of the variance in the criterion variable (CARCS) with WPCT total score accounting for 16% of that variance. If NTE score is used with the other variables, the equation accounts for 33.6% of the variance in CARCS. Both of the equations are statistically significant ($p \leq .01$), and both standardized tests are significant contributors to the multiple regression.

The next step, then, was to extend the use of CARCS to the elementary level. The process was accomplished through the use of trained classroom observers. Secondary teachers were also observed during the second phase of the study. Table 3 presents the results of those analyses.

See Table 3

Analysis of Table 3 suggests that the clarity variable exhibited the only consistent relationship with selection variables across both school levels. Apparently, the WPCT scores were good contributors in the prediction of teachers' subsequent behavior regarding clear speech and correct Standard English usage. In addition, it appears that the same scores may have had differential predictive power across school level. They were significant contributors in predicting whether or not secondary teachers subsequently used examples, used stepwise instruction, were understanding, and presented information logically, but they offered no contribution in predicting the same behaviors for elementary school teachers.

Table 3

Correlations between Certain Observation Cluster Scores and Selection Variables for Elementary and Secondary Teachers

Selection Variable	(N)	Observation Clusters[a] 1	2	3	4	5	6
Elementary							
WPCT-Verbal	(57)	-.01	.01	-.02	-.08	-.00	.32**
WPCT-Math	(57)	-.00	.05	-.00	-.04	-.04	.23*
WPCT-Total	(57)	-.01	.03	-.01	-.08	-.02	.33**
Interview	(57)	.14	.05	.12	.00	.06	.07
Essays[b]	(20)	-.23	-.30	-.06	.25	-.25	.08
Secondary							
WPCT-Verbal	(50)	.24*	.19	.26*	.09	.21	.45***
WPCT-Math	(50)	.40**	.33**	.23	.06	.26*	.34**
WPCT-Total	(50)	.34**	.27*	.27*	.08	.25*	.44***
Interview	(50)	.08	.21	.27*	.25*	.22	.18
Essays[b]	(21)	-.04	-.13	-.00	-.00	-.03	.24

[a] 1 - Elaboration-Integration, 2 - Mnemonic-Differentiation
3 - Integrity-Rapport, 4 - Application-Generalization, 5 - Information Delivery, 6 - Clarity

[b] Rank Correlations used with this variable

*p $\leq$ .05 level of significance
**p $\leq$ .01 level of significance
***p $\leq$ .001 level of significance

It should be noted that the coefficients of determination may be underestimates of the relationship between true scores on any of the paired variables. For example, if a correction for errors of measurement were used, the coefficient of correlation between the Clarity variable and the WPCT-Verbal score would be approximately .60 —— a substantial increase over the .45 calculated on observed scores. The same correction coefficient would indicate the upper limit of the relationship between any two sets of scores on the two particular variables.

When multiple regression equations were computed on the data, the shared variance of the Clarity variable with the WPCT-Verbal score was both significant and substantial ($p \leq .001$, beta = .50, 25% shared variance for secondary school teachers; $p \leq .01$, beta = .40, 16% shared variance for elementary school teachers). Overall results of the multiple regression of WPCT scores on combined cluster ratings are presented in Table 4.

See Table 4

The multiple correlation coefficients suggest a strong relationship between WPCT and quality of instruction as measured by classroom observations of elementary and secondary school teachers and by student ratings of secondary school teachers. The relationship appears to be stronger for secondary teachers. It is important to note that neither the structured interview nor the scored essay predicted classroom performance nearly as well. A second series of studies, to be discussed shortly, examined this problem.

Table 4

Multiple Regression Coefficients for PCT Scores on the Three Sets of Combined Cluster Ratings of Quality of Instruction

		Wesman PCT Scores					
		WPCT-Verbal		WPCT-Verbal		WPCT-Total	
Combined Clusters	(N)	Mult-R	R^2	Mult-R	R^2	Mult-R	R^2
Clusters from Elementary School Observations	(57)	.50	.25	.52*	.27	.57*	.32
Clusters from Secondary School Observations	(50)	.70***	.49	.78***	.61	.76***	.58
Clusters from Student Rating Scale	(49)	.56*	.31	.65**	.42	.61**	.37

*p < .05 level of significance
**p < .01 level of significance
***p < .001 level of significance

The final criterion that should be used to determine an effective teacher is the achievement of the students that have been taught by that teacher. Unfortunately, criterion-referenced tests were not available to use as student outcome measures (a shortcoming that will soon be remedied). Standardized tests are often too general to be used in a manner so precise as to attempt to determine individual student gains from a particular period of instruction. However, since standardized test scores were available, they were used with the full knowledge that they probably would not be sensitive enough to measure short-term achievement gains that could be attributed to teacher influence.

For the measure of student achievement, pretest (Spring, 1978) and posttest (Spring, 1979) scores on each subtest of the ITBS and the ITED were extracted from data-base files. Students were matched to teachers, classroom means were computed, and residual mean gain scores were obtained for each classroom at the elementary school level. Rank correlations were then calculated between residual mean gain scores and selection variables.

At the secondary school level, the number of classrooms per analysis was greatly reduced due to specialization. That is, only language arts teachers were used in the language arts analyses, math teachers in the math analyses, etc. Due to the small sample sizes, residual mean gain scores were not computed. Rather, mean gain scores were computed for each classroom.

The results were encouraging, if not overwhelming. At the elementary level, WPCT scores did not predict student achievement to any significant degree. At the middle school level (grades 7-8), results were much more substantial. The WPCT predicted math achievement consistently and well (Spearman rank correlations between WPCT-Total and Math Concepts was .78, between WPCT-Total and Math Total .61; $p \leq .01$

and .05, respectively). The most outstanding statistic here was the ability of the WPCT-Verbal subtest to predict student achievement on the Math Concepts subtest of the ITBS (r_s=.84, $p \leq .01$). At the senior high school level, WPCT scores from one subtest or the other predicted performance in reading (r_s=.57, $p \leq .05$), language arts-usage (r_s = .61, $p \leq .05$), language arts-spelling (r_s = .54, $p \leq .05$), language arts-total (r_s = .54, $p \leq .05$), language arts-total (r_s = .67, $p \leq .05$), social studies (r_s = .87, $p \leq .05$), and science (r_s = .70, $p \leq .05$).

Thus, even though standardized achievement tests were used as the criterion measure, teachers' performances on the WPCT were significantly related to their students' performances on the ITED at the secondary level in seven of twenty-seven comparisons and yielded substantial positive correlations ($r_s \geq 30$) in another seventeen comparisons (N's varied from 5 to 11). At the middle school level, significant correlations occurred between teachers' performances on the WPCT and their students' performances on the ITBS in six of twelve comparisons with substantial positive correlations ($r_s \geq .30$) in another three (N's varied from 7 to 10). No significant relationships were found at the elementary level. Overall, these results across school levels yielded four times as many significant results as would be expected by chance.[8]

---

[8] When it became necessary to replace the WPCT, a similar instrument, the Personnel Services Test (PST) was developed. The criterion-related (concurrent) validity of the PST-Verbal was determined by computing the Pearson product-moment coefficients of correlation (r) between PST-Verbal and Wesman PCT-Verbal scores. Construct validity of the PST was investigated using two approaches: Computing Pearson r's between PST-Verbal scores and the two Personnel Services ratings that relate indirectly to the same theoretical construct, (presumably, "educational training" and "educational background" are ratings that would be expected to correlate at least moderately with a measure of "breadth of academic knowledge."), and factor analyzing of test items. Content validity of the PST was determined on the basis of an evaluation of the processes of item generation and item selection. Details of the validation of the PST are specified in Carter and Taite (SP80-401-31-05).

It is believed that the failure of the WPCT to predict elementary school classroom achievement is as much the fault of the ITBS as it is of the WPCT. The ITBS is a very good norm-referenced standardized test. As such, it is a survey test. Survey tests are not designed to be sensitive to relatively short periods of instruction. In addition, and as a consequence, the variance in student test scores and subsequent gain scores are not as great, as measured by standardized achievement tests, at the elementary level as they are at the secondary level. That is, test score variance increases as grade level increases. The WPCT would probably have much greater success predicting elementary achievement on more focused criterion-referenced tests. This hypothesis will be tested as criterion-referenced tests become increasingly available in the DISD elementary grades. The secondary level was less affected by the aforementioned phenomenon because of the increased variance in test scores and subsequent gain scores.

Relatively, WPCT scores were more powerful predictors of classroom performance than were interview scores or essay scores. The relative ineffectiveness of interview and essay scores in predicting classroom performance suggested a need to refine the instrumentation and scoring procedures further on those exercises. The complete lack of relationship between principal's evaluation of teachers and anything of consequence suggests a need for the overhaul of most teacher evaluation systems as we know them.

Prior to 1980, the major preponderance of work had been done upon ensuring the validity (construct, content and predictive) of the tests used in the system. This effectively meant that the essay and interview processes were not contributing reliable information to the

screening process.[9] In 1980, both of these problems were addressed.

The essay exercise was substantially revised. In fact, a completely new instrument and procedure were developed and implemented.[10] The content validity of the exercises was the paramount consideration in their development. The purpose of this component of the applicant screening system was to gain valid information about each examinee's ability to deal with three specific components of writing. Those writing components were: (1) legibility of handwriting; (2) mechanical skills - punctuation, grammar, capitalization, and spelling; and, (3) composition - a composite rating of clarity, congruence, and organization.

The essay exercise was tried on 156 applicants. Raters were carefully trained and provided with a scoring guide precisely defining the variables of interest. The estimates of inter-rater reliability for the initial trials were .66 for handwriting, .68 for composition, and .80 for mechanical skills. Work was done to decrease the choice of topics available to the applicant which subsequently decreased topic effect and increased the reliability of the writing sample. The current system has a choice of five topics. The raters also currently have the benefit of a revised and updated scoring guide. As a result of this work, the essay exercise has become more reliable and, thus, has the potential to contribute more valid information to the selection process.

---

[9] While both were used extensively in the applicant screening process, data reported earlier in this paper suggest that neither contributed much reliable information.

[10] Details of the development and validation of the Personnel Services Departmental Essay Test (PSDET) are specified in Taite (SP80-404-31-05).

The interview process has traditionally contributed substantially to the decision to hire or not to hire an applicant. Despite the importance of the procedure, the District was unable to present any substantive evidence as to its validity. Eight subscales were developed to be utilized in the interview process. These eight subscales were general appearance, clarity of speech, emotional stability, educational background, community relations, personality traits, goals and philosophy, and ability to communicate.[11] The interview instrument was standardized and results correlated with other similar criteria. Table 5 depicts the relationship between the eight interview subscale measures and other selection criteria.

See Table 5

Study of Table 5 suggests that there were a number of significant correlations. It is important to note that the ones that logically should have been significant, were indeed so. For example, clarity of speech correlated with PCT-Verbal, PCT-Math, and educational training; educational background correlated with the aforementioned three variables plus experience; and, ability to communicate correlated with PCT-Verbal, PCT-Math, and educational training. There were no examples of correlations not occurring that logically should have. This was interpreted as evidence for the concurrent validity of the interview process.

---

[11] Details about the revised interview process are specified in Carter and Berry (SP80-403-31-05).

Table 5

Pearson Correlations between the Eight Interview Subscale Measures and Other Selection Criteria
(n = 119)

Interview Subscale	Selection Criteria Other than Interview						
	Appli-cation	Educ. Train.	Exper-ience	Refer-ence	PCT-Verbal	PCT-Math	Essay
General Appearance	.06	.08	-.14	.05	.22**	.23**	.10
Clarity of Speech	.18*	.29***	.06	.17*	.32**	.31***	.11
Emotional Stability	.13	.28***	.01	.06	.25**	.24**	.17*
Educational Background	.18	.50***	.27**	.13	.40***	.31***	.11
Community Relations	.00	.11	-.10	.10	.10	.15	.03
Personality Traits	.19*	.33***	.13	.19*	.18*	.27**	.09
Goals and Philosophy	.12	.28***	-.03	.09	.15	.23**	.04
Ability to Communicate	.21*	.39***	.01	.20*	.31***	.29***	.21*

*p < .05
**p < .01
***p < .001

The most important measure of reliability on a rating scale that is utilized by more than one observer is an estimate of interrater consistency. Since teacher applicants are interviewed by one of seven interviewers, scoring strategies and standards must be consistent across raters. Seven of the regularly assigned District interviewers viewed videotapes of eighteen teacher applicants and rated them using the procedures utilized in the official interview. These videotapes were later used in interviewer training.

Reliability of observation or interview procedures is influenced by several forms of rater errors, one of the most common of which is the error of leniency. This type of error refers to the tendency on the part of some raters to give higher (or lower) scores than others. Since an applicant is interviewed by only one person, the error due to leniency, if present, could be a problem for obvious reasons (the advantage of being interviewed by a rater who gives high marks and the disadvantage of being interviewed by one who tends to give low marks are not equitable situations).

Table 6 presents the average ratings of all 18 teacher applicants for each interviewer. The table indicates an absence of leniency effects when the group of applicants is considered at once. The only exceptions are in the Emotional Stability and Community Relations subscales for interviewers F and G, respectively. The important comparison between raters is in the average total scores, and here there is much congruence among interviews (largest difference is 2.2 points on a 50-point scale).

See Table 6

Table 6

Average Interview Scores of Videotaped Teacher Applicants Rated by Seven Interviewers

Subscale	(Total Points)	Interviewer A	B	C	D	E	F	G
General Appearance	(4)	3.3	2.8	3.3	2.8	2.8	2.9	3.2
Clarity of Speech	(6)	4.2	4.1	4.3	4.2	4.2	3.8	4.3
Emotional Stability	(6)	4.3	4.2	4.3	4.2	3.9	3.5*	4.4
Educational Background	(10)	6.8	6.7	7.1	6.8	6.7	6.7	7.1
Community Relations	(2)	1.1	1.4	1.3	1.2	1.1	1.6	0.8*
Personality Traits	(10)	6.6	6.4	6.4	6.6	6.3	6.4	6.9
Goals and Philosophy	(4)	2.6	3.0	3.2	2.7	2.8	3.0	2.8
Ability to Communicate	(8)	5.8	5.9	6.1	5.9	5.8	5.6	5.8
Total	(50)	34.7	35.9	35.9	35.0	33.7	36.0	35.3

*Exceptions

How consistent the same raters were when they gave scores for particular individuals and how closely they estimated true scores for those individuals are addressed in Table 7. A conventional rule-of-thumb is to regard a reliability estimate of .50 or greater for a single observation and .80 or greater for group observations as acceptable. Five of the eight subscales do not meet one or both of these criteria. Particularly noticeable are the low estimates for Emotional Stability, Community Relations, and Goals and Philosophy. This may not be a nagging problem, however, since it is only the total interview score that is given the preponderance of attention when screening applicants. The estimates for this score indicate good reliability.

See Table 7

The Average Point Differential refers to the average range in scores among raters for each interviewee on each subscale. The six-point average differential on the total scores suggests that, on the average, at least two raters disagreed by a magnitude of over six points on scores given each interviewee. Since the reliability estimates are rather substantial, however, this average may be somewhat misleading. That is, the majority of interviewers were in apparent agreement on total scores as indicated by the high reliability estimate (.93).

Generally, the reliability of interview scores is good (and has probably increased with the ability to train interviewers through the use of the videotapes). This is particularly true for the total interview score, which is the measure which is given almost exclusive

Table 7

Reliability Estimates of Interview Scores
of Videotaped Teacher Applicants

Subscale	(Total Points)	Reliability Estimate of Single Score	Reliability Estimate of Group Score	Average Point Differential
General Appearance	(4)	.48	.87	.89
Clarity of Speech	(6)	.52	.88	1.28
Emotional Stability	(6)	.35	.79	1.61
Educational Background	(10)	.50	.87	1.56
Community Relations	(2)	.31	.76	.94
Personality Traits	(10)	.54	.89	1.83
Goals and Philosophy	(4)	.37	.80	1.28
Ability to Communicate	(8)	.43	.84	1.61
Total	(50)	.66	.93	6.33

consideration when evaluating an applicant's potential success from the interview. There appear to be few rater errors due to leniency effects and a high level of correspondence among interviewers on total scores averaged across all applicants in the study.

Table 8 depicts the relationship among the various variables used in the teacher selection system after the essay and interview procedures had been revised. These data were, of course, collected from a new sample of applicants. The data in Table 8 compare very favorably with the prerevision statistics cited in Tables 1, 2, and 3.

See Table 8

In summary, a series of validation studies, briefly summarized in the paper and conducted over a four year period, provided evidence for the predictive, concurrent, construct, and/or content validity of various components of the teacher selection system. Although additional studies need to be conducted, the most important of which is a cross-validation study designed along the lines of the first two studies reported in the paper, it appears that the DISD has the instrumentation and procedures to select teachers reliably and validly from the annual applicant group. The relevant question is, how has the process worked?

A Retrospective Analysis

The work completed through the series of studies reported in the paper suggests that verbal and quantitative ability, as measured by the WPCT and the subsequent PST, are good predictors of teacher success in the classroom. It is important to note that the major purpose of a

Table 8

Pearson Coefficients of Correlation among Selection Criteria on A Teacher Applicant Sample (N = 119)

	WPCT-Math	Interview	Application	Training	Experience	References	Essay
WPCT-Verbal	.57***	.35***	.01	.24**	.04	.17	.29***
WPCT-Math		.36***	.06	.26**	.03	.10	.31***
Interview			.21*	.43***	.09	.18*	.16*
Application				.22**	.10	.09	.07
Educational Training					.18*	.26**	.18*
Experience						.05	-.02
References							.14

*p < .05
**p < .01
***p < .001

preemployment assessment system is efficiently to select teachers who have a high probability of success in the classroom. Since there are many more applicants than the District is able to hire, efficiency of selection is paramount. The purpose is much different from that of a personnel evaluation system where the goal is, presumably, improvement of the existing teaching force. As a result, within the context of a screening and selection system, short tests of verbal and quantitative ability appear to be good indicators of potential teacher performance.

The real world of education, however, tends not to emphasize this relationship in the selection of teachers. A recent survey (Taite and Carter, 1980) reported that, nationally, the interview process appears to carry the most weight in the selection of teachers. Aware of the national trend, and concerned that the tests of verbal and quantitative ability were being relied upon too heavily in the selection of teachers, an investigation was conducted on the relative weights that various selection criteria had in the selection procedure.

Table 9 presents the results of a discriminant analysis of teacher selection data. These data were collected from a sample of hired versus non-hired teacher applicants. The discriminant analysis provides weights to each of the relevant variables in the screening process such that their linear combination provides the greatest separation between the group of teacher applicants who were selected versus the group who were not selected.

See Table 9

Study of Table 9 suggests that the result of the interview is the

Table 9

Discriminant Function Coefficients (Weights)
for Teacher Selection Variables in
Predicting Employment
(N = 119)

Selection Variable	Weight	Selection Variable	Weight
Essay Test	-.20	Application	.16
PST-Verbal	.15*	Educational Training	.13**
PST-Math	.03	Experience	.08
Interview	.56***	References	.52***

Note. Significance levels are based upon one-way analyses of variance F-tests for equality of means for the hired and non-hired groups on a single selection variable.

*p < .05
**p < .01
***p < .001

single most important predictor in the hiring process. Personal references are the second most important predictor. The PST-Verbal test score and educational training also contribute significantly to the process (although not nearly so heavily weighted as the interview process and personal references).

It appears that, at least in terms of the amount of importance attributed to the interview process, the DISD is consistent with the national trend. The difference is that, in this case, much attention has been devoted to studying and improving the reliability, as well as the validity, of the instrument. It also appears that the District need not worry about the PST-Verbal and PST-Math scores being relied upon too heavily in the applicant screening process. If anything, they probably do not contribute enough. The negative weight assigned to the essay examination is probably the result of differential emphasis on this component, depending on the teaching field, and the fact that the DISD has been concentrating on hiring math and science teachers.

Formal studies of the applicant screening process were discontinued in 1981. At this point, the District is awaiting the completion of its criterion-referenced testing system (scheduled for Spring, 1985) before pursuing further validation studies. When the criterion-referenced testing system is complete, studies similar to the first two outlined in the paper will again be conducted. The student achievement variable will be measured through the use of criterion-referenced tests rather than norm-referenced tests. These cross-validation studies should, as a result, provide a more precise measurement of the student achievement variable. The reliability of the essay exam will be re-examined and the reliability of the interview process again carefully scrutinized.

It is difficult to assess the overall impact of the teacher selection system. The applicant screening process has been studied, overhauled, and significantly improved. Objective measures of applicant ability in basic skills areas have been utilized in the hiring process. DISD principals and other administrators are convinced that teachers hired under the current system are, on the whole, superior to teachers hired under the previous system. During the period of time that this system has been in operation, student achievement scores have experienced unprecedented increases. Although many educational reforms have been undertaken, most participants in the process attribute much of the success to better teachers. This process has contributed substantially to the employment of better teachers.[12]

---

[12] The State of Texas has subsequently adopted an applicant screening program for prospective teachers.

REFERENCES

Carter, Wallace H. An Interpretive Analysis Of The Personnel Services Research Studies, 1980-81, CS81-804-61-05, DISD, December, 1981.

Carter, Wallace H. An Interpretive Analysis Of The Teacher Selection Validity Studies, RE80-804-61-05, DISD, October, 1980.

Carter, Wallace H., and Berry, Renita. Evaluation Of The Interview Process For Teacher Applicants, SP80-403-31-05, DISD, October, 1980.

Carter, Wallace H., Casey, Kathleen, and Taite, Glenna. Teacher System Validation Study, IR78-283-31-12, DISD, September, 1978.

Carter, Wallace H., Casey, Kathleen, and Taite, Glenna. Wesman Personnel Classification Inventory Norming Study, IR78-236-31-05, DISD, December, 1977.

Carter, Wallace H., and Taite, Glenna. Teacher Selection Validity Study, AA79-100-31-05, DISD, September, 1979.

Carter, Wallace H., and Taite, Glenna. Validation Of The Personnel Services Test, SP80-401-31-05, DISD, October, 1980.

Feistritzer, C. Emily. The Condition of Teaching: A State By State Analysis, A Carnegie Foundation Technical Report, Princeton University Press, 1983.

Ford, D. L., et. al. Perceptions, Opinions, and Attitudes Toward The Dallas Independent School District and School Desegregation, December, 1979.

Gage, N. L., ed. Handbook of Research on Teaching, Chicago: Rand McNally, 1963.

Gagne, Robert M. and Briggs, Leslie J. Principles of Instructional Design, New York: Holt, Rinehart and Winston, 1974.

Gallup Poll of Public Attitudes Toward Education, 1979.

Levin, Henry. "A Cost-Effectiveness Analysis of Teacher Selection," Journal of Human Resources 5:24-33, Winter, 1970.

Lyons, Gene. "Why Teachers Can't Teach," Texas Monthly, September, 1978, pp. 123-129, 208-220.

Northern, E. F. "The Trend Toward Competency Testing of Teachers," Phi Delta Kappan, January, 1980, p. 359.

Taite, Glenna L. The Development Of The Personnel Services Department Essay Test, SP80-404-31-05, DISD, October, 1980.

Taite, Glenna and Carter, Wallace H. "The Use of Written Examinations and Interviews For Teacher Selection In Large Urban School Districts," Telephone Survey, DISD, January, 1980.

Travers, Robert M. W., ed. Second Handbook of Research on Teaching, Chicago: Rand McNally, 1973.

U. S. News and World Report "America's Teachers - Are They To Blame?," September 11, 1978, pp. 53-54.

U. S. News and World Report "Quest For Better Schools," September 11, 1978, pp. 50-52.

United States vs. State of South Carolina, District Court: 15 FEP Cases 1196 (1977); Supreme Court: 38 CCH S. Ct. Bull. p. B559 (1978).

Vlaanderen, Russ. Trends In Competency-Based Teacher Certification, Education Commission of the States, May, 1979.

Webster, William J., "The Validation of a Teacher Selection System," AERA, Boston, 1980.